回归分析在数模竞赛中的应用

这一讲义已经写成 Word 文件，放在一个网盘中，可以按照下列步骤下载：

（1）在浏览器地址栏中打入“ http://mail.163.com/ ”，进入网易 163 邮箱。

（2）写入用户名“mathcai_lu”和密码“111111”，点击“登录网盘”按钮，进入网盘。

（3）点击“数学模型”，进入“数学模型”文件夹。

（4）在“回归分析在数模竞赛中的应用.exe”文件名前面打勾，在上面的操作栏中点击“下载文件”的图标。在弹出窗口中，点击“保存”按钮，就可以将“回归分析在数模竞赛中的应用.exe”文件下载到你的硬盘上。

（5）在你的硬盘上运行“回归分析在数模竞赛中的应用.exe”文件，就会在你的硬盘上自动释放生成 4 个名为“回归分析在数模竞赛中的应用-?.doc”写有讲义的文件

§1 回归分析的基本思想

在实际问题中，我们会遇到各种变量，在变量与变量之间，往往存在着各种关系。

例如，圆的半径 R 与圆面积 S 之间，有这样的关系：$S=\pi R^2$。

又如，自由落体落下的时间 t 与落体落下的距离 h 之间，有这样的关系：$h=\frac{1}{2}gt^2$。

……，等等。

在这些关系中，只要自变量的值确定了，因变量的值也就随之确定了。像这样的变量之间的关系，是**确定的函数关系**。

但是，有些变量之间的关系就不是这样。

例如，农作物的施肥量 x 与农作物的产量 y 之间的关系。

又如，商品的价格 x 与商品的销售量 y 之间的关系。

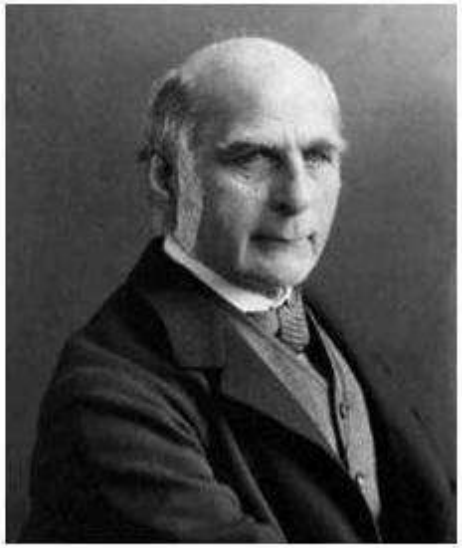
又如，家庭的收入 x 与家庭的支出 y 之间的关系。

又如，父亲的身高 x 与儿子的身高 y 之间的关系。

……，等等。

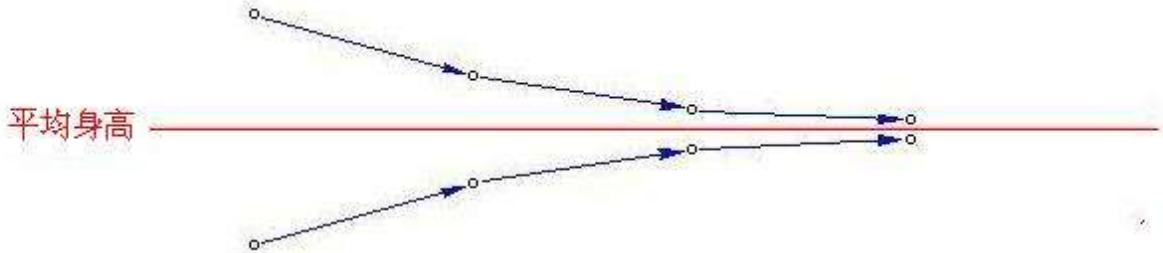
在这些关系中，自变量 x 的值确定了，因变量 y 的值并不完全随之确定，还是可能有上下起伏的变化。时，在这些关系中，自变量 x 与因变量 y 又不是完全无关的，通过大量的统计数据，可以发现，它们之间确存在着某种关系。我们把这样的关系，称为**统计相关关系**。

回归分析，就是研究变量之间的统计相关关系的一种统计方法。



高尔顿（ Francis Galton，1822—1911）英国人，达尔文的表兄弟，从小聪颖过人。高尔顿原来与达尔文一也是研究生物学的，但是，后来他的兴趣逐渐转向人类学、尤其是遗传优生学的研究。他担任了英国伦敦大学授，首次开设了遗传优生学讲座，还创办了一个优生学实验室，他被认为是现代遗传优生学的创始人。

高尔顿在研究人的身高遗传时，发现了一个有趣的现象。



1886 年，高尔顿发表论文《遗传中向平均身高回归的现象》，引起了在伦敦大学担任数学和力学教师的卡尔尔逊的注意，两人决定合作研究这一现象。

卡尔·皮尔逊（Karl Pearson，1857-1936）英国人，1901 年，与高尔顿一起，创办了《Biometrika（生物计量杂志，是现代数理统计学的创始人。

他们收集了 1078 对父亲和儿子身高的数据：

父身高 子身高

(x_i, y_i)，$i = 1, 2, \cdots, 1078$

得到直线的方程为

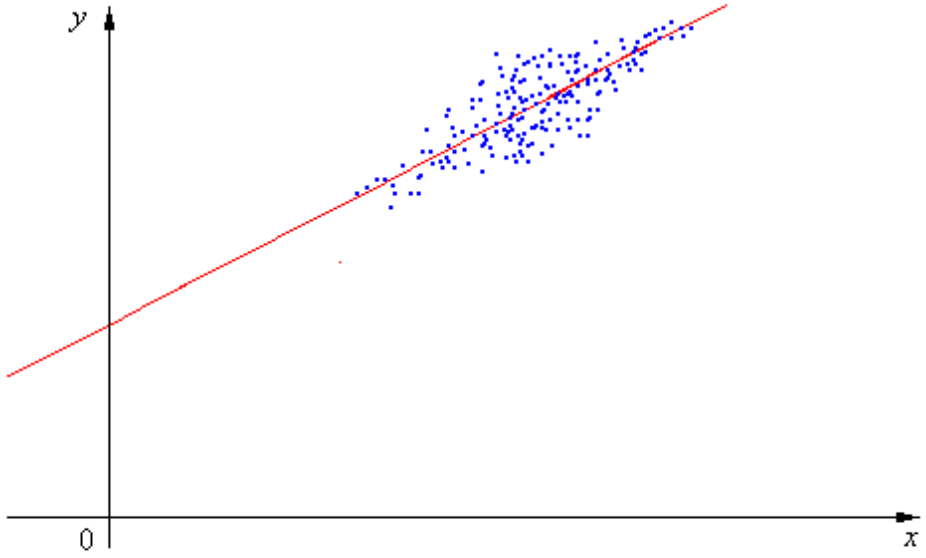
$\hat{y} = 0.8567 + 0.516x$ （单位：米）

例：$x = 1.900 \rightarrow \hat{y} = 1.837$

$x = 1.837 \rightarrow \hat{y} = 1.805$

$x = 1.600 \rightarrow \hat{y} = 1.682$

$x = 1.682 \rightarrow \hat{y} = 1.725$



回归分析（Regression Analysis）——从自变量和因变量的一组观测数据出发，寻找一个函数式，将变量之
统计相关关系近似地表达出来。这个能够近似表达自变量与因变量之间关系的函数式，称为**回归方程**（Regr
Equation）或**回归函数**（Regression Function）。

§2 回归分析问题的一般形式

设有 m 个自变量 $x_1, x_2, \cdots, x_m$ 和 1 个因变量 y，它们之间有下列关系：

$$y = F(x_1, x_2, \cdots, x_m; a_1, a_2, \cdots, a_p) + \varepsilon \quad ,$$

其中，F 是函数形式已知的 m 元函数，$a_1, a_2, \cdots, a_p$ 是常数，是函数 F 中的未知参数，ε 是表示误差的随机变量，一般可认为 $\varepsilon \sim N(0, \sigma^2)$，$\sigma > 0$ 。

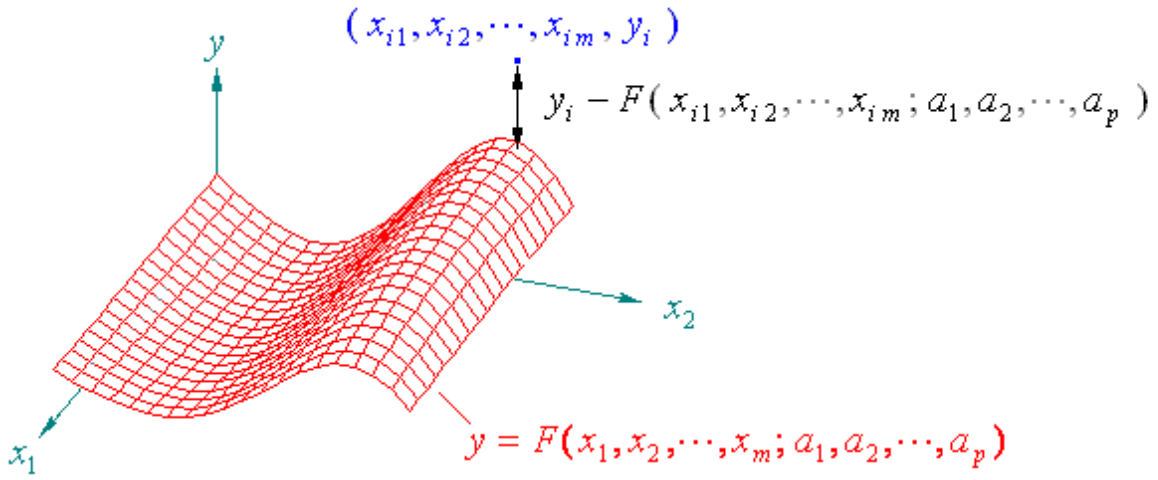
对 $x_1, x_2, \cdots, x_m$，y 进行 n 次观测，得到观测值：

$$(x_{i1}, x_{i2}, \cdots, x_{im}, y_i) \quad , \quad i = 1, 2, \cdots, n \quad 。$$

对每一次观测来说，同样有下列关系

$$y_i = F(x_{i1}, x_{i2}, \cdots, x_{im}; a_1, a_2, \cdots, a_p) + \varepsilon_i \quad ,$$

其中 ε_i 是第 i 次观测时的随机误差，$i = 1, 2, \cdots, n$ 。



回归分析目标是：从观测数据出发，求 $a_1, a_2, \cdots, a_p$ 的估计 $\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p$ ，使得下列平方和 Q 达小：

$$Q = \sum_{i=1}^{n} [\, y_i - F(x_{i1}, x_{i2}, \cdots, x_{im}; a_1, a_2, \cdots, a_p)\,]^2 \quad 。$$

由于估计的目标是使一个平方和达到最小，而平方又称为“二乘”，所以，这种估计称为**最小二乘估计**（Squares Estimator，**简称** LSE ），求这种估计的方法称为**最小二乘法**（ Method of Least Squares ）。

把最小二乘估计 $\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p$ 代入 Q 表达式，就得到 Q 的最小值

$$Q_{\min} = \sum_{i=1}^{n} [\, y_i - F(x_{i1}, x_{i2}, \cdots, x_{im}; \hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p)\,]^2 \quad 。$$

Q 的最小值称为**残差平方和**（也称**剩余平方和** Residual Sum of Squares ，简称 RSS ），残差平方和越小，说归方程表达变量之间统计相关关系的精确程度越高，也就是回归分析的效果越好。

在数模竞赛中，经常会遇到可以用回归分析来解决的问题，下面是一些例子。

例 1（1993 年全国数模竞赛 A 题）非线性交调的频率设计

在一个电子通讯系统中，对输入信号强度 u 和输出信号强度 y 进行观测，得到下列数据：

u	0	5	10	20	30	40	50	60	80
y	0	2.25	6.80	20.15	35.70	56.40	75.10	87.85	98.50

已知 u 与 y 之间的关系，是一个次数为 3 次的多项式：

$$y = \beta_0 + \beta_1 u + \beta_2 u^2 + \beta_3 u^3 + \varepsilon \ ,$$

作为非线性交调的频率设计的第一步，需要求出这个关系式。

这里，u 是自变量，y 是因变量，$\beta_0, \beta_1, \beta_2, \beta_3$ 是未知参数。问题是要从 u 和 y 的观测值数据出发，

求出参数 $\beta_0, \beta_1, \beta_2, \beta_3$ 的估计。显然，这是一个回归分析问题。

例 2（1993 年国际数模竞赛 A 题）加速餐厅剩菜堆肥的生成

一家自助餐厅，每天把顾客吃剩下的食物搅拌成浆状，混入厨房里废弃的碎绿叶菜和少量撕碎的报纸，再加
真菌和细菌，混合物原料在真菌和细菌的消化作用下生成堆肥。

下表给出了以磅为单位的混合物原料中各种成分的的数据，以及混合物原料喂入的日期和堆肥生成的日期

食物浆	绿叶菜	纸片	原料喂入日期	堆肥生成日期
86	31	0	90.7.13	90.8.10
112	79	0	90.7.17	90.8.13
71	21	0	90.7.24	90.8.20
03	82	0	90.7.27	90.8.22
79	28	0	90.8.10	90.9.12
105	52	0	90.8.13	90.9.18
121	15	0	90.8.20	90.9.24
110	32	0	90.8.22	90.8.22
82	44	9	91.4.30	91.6.18
57	60	6	91.5. 2	91.6.20
77	51	7	91.5. 7	91.6.25
52	38	6	91.5.10	91.6.28

要求确定：混合物原料中各种成分的比例与堆肥生成的速率之间是否有关系？如果有关系，怎样的比例才能
得堆肥生成的速度最快?

设 x_1,x_2,x_3 分别是食物浆、绿叶菜和纸片在混合物原料中的比例，y 是生成堆肥所需要的时间。要尝

出 x_1,x_2,x_3 与 y 之间的关系式。可以考虑各种不同形式的关系，最简单的，可以认为它们之间有线性关系

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon\,,$$

其中，x_1,x_2,x_3 是自变量，y 是因变量，$\beta_0,\beta_1,\beta_2,\beta_3$ 是未知参数。

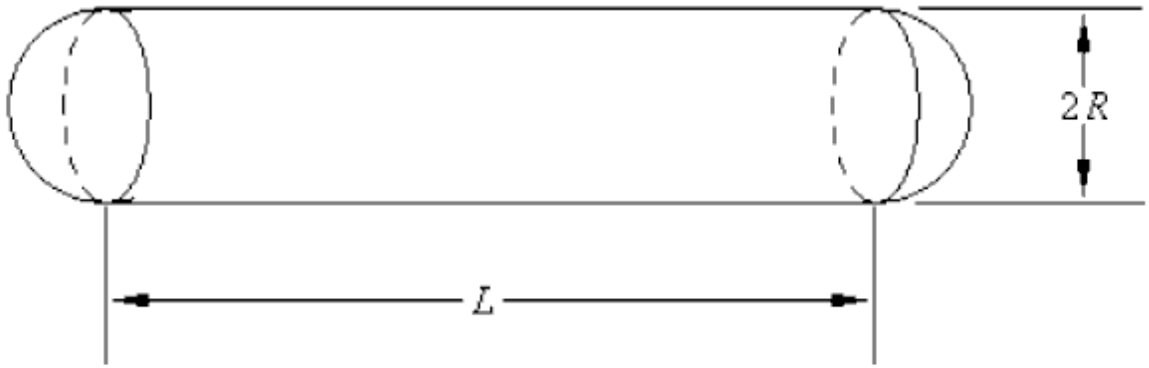
问题是要从 x_1,x_2,x_3 和 y 的观测值数据出发，求出参数 $\beta_0,\beta_1,\beta_2,\beta_3$ 的估计（由于 x_1,x_2,x_3 是

成分在总量中的比例，它们之间有 $x_1+x_2+x_3=1$ 的关系，3 个自变量实际上不是独立的，为了避免估计结

不确定，实际上还应该去掉一个自变量）。显然，这也是一个典型的回归分析问题。

例 3（1996 年国际数模竞赛 A 题）潜水艇的探测

海洋中有一个背景噪声场，当附近有潜水艇驶过时，噪声场会发生变化。要求给出一种方法，通过在水下点检测到的噪声场的变化情况，探测出附近有无潜水艇，潜水艇的位置、大小、形状、运动速度和运动方向。



设 (x_0, y_0, z_0) 是潜水艇中心的坐标，(V_x, V_y, V_z) 是潜水艇的速度分量。近似认为潜水艇的形状是一个形的主体，前后两端加上两个半球。设 L 是潜水艇圆柱形主体的长度，R 是圆柱形底面的半径。

在海洋中设置 n 个检测点。设第 i 个检测点的坐标位置为 (x_i, y_i, z_i)，在这一点上测到的噪声强度为 $i = 1, 2, \cdots, n$。根据水声学原理，可以得到下列形式的关系式：

$$p_i = F(x_i, y_i, z_i; x_0, y_0, z_0, V_x, V_y, V_z, L, R) + \varepsilon_i \quad , \quad i = 1, 2, \cdots, n \text{。}$$

问题是要从自变量的观测值数据 x_i, y_i, z_i 和因变量的观测值数据 p_i 出发，求出未知参数 $x_0, y_0, z_0, V_x, V_y, L, R$ 的估计。显然，这也是一个回归分析问题。

§3 线性回归（Linear Regression）

一、线性回归问题的一般形式和解法

设有 m 个自变量 $x_1, x_2, \cdots, x_m$ 和 1 个因变量 y，它们之间有下列关系：

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \quad ,$$

其中，$\beta_0, \beta_1, \cdots, \beta_m$ 是未知参数，$\varepsilon \sim N(0, \sigma^2)$ 是表示误差的随机变量，$\sigma > 0$ 。

对 $x_1, x_2, \cdots, x_m$，y 进行 n 次观测，得到一组观测值：$(x_{i1}, x_{i2}, \cdots, x_{im}, y_i)$ ，$i = 1, 2, \cdots, n$ 。

即有

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \quad , \quad i = 1, 2, \cdots, n \quad 。$$

要求从自变量和因变量的观测数据出发，求未知参数 $\beta_0, \beta_1, \cdots, \beta_m$ 的估计值 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_m$ ，使得

$$Q = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im})]^2$$

达到最小。

Q 是 $\beta_0, \beta_1, \cdots, \beta_m$ 的函数，所以，这是一个多元函数求最小值的问题，我们可以通过求偏导数、解

方程组的方法，来确定 Q 的最小值点：

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = 0 \\ \cdots\cdots \\ \dfrac{\partial Q}{\partial \beta_m} = 0 \end{cases}$$

从这个方程组中求得的解 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_m$，使 Q 达到最小，是 $\beta_0, \beta_1, \cdots, \beta_m$ 的最小二乘估计。（有时

线性回归问题中可能会不出现常数项 β_0，也可以类似地求解。）

当自变量个数 n 比较多时，线性回归的具体计算是很烦琐复杂的，如果靠人工计算，工作量很大。现在计

机已经十分普及，人们已开发了许多现成的计算机程序和软件包，其中包括可以作一元和多元线性回归的软件。

们在解决实际问题时，可以利用这些现成软件，十分方便迅速地完成线性回归的计算。所以，我们这里就不将线

回归的具体计算公式详细写出来了。

二、衡量线性回归结果好坏的标准

（1）残差平方和（剩余平方和 Residual Sum of Squares ，简称 RSS），

残差平方和，也就是 Q 的最小值，记为

$$SS_e = Q_{\min} = \sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_m x_{im})]^2 \text{ 。}$$

SS_e 越小，说明回归方程表达变量之间统计相关关系的精确程度越高，也就是回归分析的效果越好。但 SS_e 大小还与样本观测次数 n 有关。

（2）估计的标准差（残差标准差 Estimated Standard Deviation）

$$\hat{\sigma}_e = \sqrt{\frac{SS_e}{n-m-1}}$$ （如果回归问题中不出现常数项 β_0，则式中的 $n-m-1$ 要改为 $n-m$ ）。

$\hat{\sigma}_e$ 越小，表明 SS_e 越小，回归分析的效果也就越好。$\hat{\sigma}_e$ 的大小基本上与样本观测次数 n 无关，但它个有量纲的量，与因变量 y 同一量纲单位，所以它的数值大小与 y 的量纲单位大小有关。

（3）多重相关系数（复相关系数 Multiple Correlation Coefficient）

$$r = \sqrt{1 - \frac{SS_e}{L_{yy}}}$$ ，其中，$L_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ ，$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ 。

可以证明，有 $0 \le r \le 1$ 。r 越接近 1 ，说明 SS_e 越小，回归分析的效果也就越好。

r 是一个无量纲的量，它的大小与量纲的单位大小无关。

三、线性回归应用的实例

例 4（1993 年全国数模竞赛 B 题）给足球队排名次

已知 12 支球队在全国甲级联赛中的成绩，要求设计一种依据这些成绩给足球队排名次的方法。

这个问题可以有多种不同的做法，回归分析就是其中的一种做法。

设 $m=12$ 支球队的实力为 $\beta_1,\beta_2,\cdots,\beta_m$ ，这些都是未知的常数。

设 y_i 是第 i 场比赛时，通过比分表现出来的主队与客队两队的实力之差。例如，当两队的比分为 $3:2$

可以定义 $y_i=3-2$ 或 $y_i=\sqrt{3}-\sqrt{2}$ 或 $y_i=\sqrt[3]{3}-\sqrt[3]{2}$ 或 $y_i=\ln\left(\frac{1+3}{1+2}\right)$ ，等等。

设第 1 场比赛，是 1 队对 2 队，1 队为主队，2 队为客队；第 2 场比赛，是 3 队对 4 队，3 队为主队，4 队

队；第 3 场比赛，是 1 队对 4 队，1 队为主队，4 队为客队；…… 。

则有

$$y_1=\beta_1-\beta_2+\varepsilon_1 \ ,$$

$$y_2=\beta_3-\beta_4+\varepsilon_2 \ ,$$

$$y_3=\beta_1-\beta_4+\varepsilon_3 \ ,$$

…… 。

对每一场比赛，有

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \varepsilon_i \quad ， \quad i = 1, 2, \cdots, n \quad 。$$

其中，

$$x_{ij} = \begin{cases} +1 & \text{第 } i \text{ 场比赛，第 } j \text{ 队作为主队参赛} \\ -1 & \text{第 } i \text{ 场比赛，第 } j \text{ 队作为客队参赛} \\ 0 & \text{第 } i \text{ 场比赛，第 } j \text{ 队没有参赛} \end{cases}$$

$\varepsilon_i \sim N(0, \sigma^2)$ 是第 i 场比赛结果的随机误差，$i = 1, 2, \cdots, n$ 。

可以看出，这实际上是一个不出现常数项 β_0 的线性回归问题，回归方程为

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon \quad 。$$

要求从观测值 $x_{i1}, x_{i2}, \cdots, x_{im}$（+1，−1 或 0）和 y_i（比赛结果）出发，求 $\beta_1, \beta_2, \cdots, \beta_m$（各队实力）的估计

求出各队实力的估计值，就可以按照实力的大小给各队排名次了。

实际计算时，还要考虑到比赛结果只反映各队的实力之差，只知道相对的大小关系，缺少一个绝对的基准想求出各队实力的数值，实际上是不可能的。要解决这个问题也很容易，只要事先给定一个球队实力的数值，一个基准就可以了。例如，可以令 $\beta_m = 0$，这相当于在回归方程中去掉最后一项，然后作线性回归，就可以其他的 β 的估计值。

§4 广义线性回归（Generalized Linear Regression）

一、广义线性回归基本思想

例 5 抛物线的拟合

某零件上有一条曲线，可以近似看作是一条抛物线，为了在数控机床上加工这一零件，在曲线上测得 n 的坐标 (x_i, y_i)，$i = 1, 2, \cdots, n$，要求从这 n 个点的坐标出发，求出曲线的函数表达式。

显然，这是一个回归分析问题，由于曲线可以近似看作是一条抛物线，因此，回归方程（即曲线的函数表达式）是一个二次多项式

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \quad 。$$

像这种回归方程是一个多项式的回归，称为**多项式回归**（Polynomial Regression）。

虽然多项式回归方程不是线性的，但可以通过变量代换，化成线性形式。

令 $x_1 = x$，$x_2 = x^2$，原来的回归方程化成了下列形式：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad ，$$

这是一个线性回归方程，可以用前面介绍过的线性回归的方法求出它的解。具体作回归时，所需要的观测值 x_{i1}，x_{i2} 用 x_i，x_i^2 的数值代入，求得的线性回归方程中常系数的估计 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$，也就是原来的二次多项式回归方程中常系数的估计。

例 6 科布-道格拉斯(Cobb-Douglas)生产函数

在经济学中，有一个著名的科布-道格拉斯生产函数，这个函数指出，生产产出 Y 与劳动投入 L 、资入 K 之间，近似有下列关系：

$$Y = \alpha L^{\beta_1} K^{\beta_2} + \varepsilon \ ,$$

其中，$\alpha,\ \beta_1, \beta_2$ 都是常系数。现测得一组劳动投入、资本投入和生产产出的数据 (L_i, K_i, Y_i)，$i = 1, 2, \cdots,$

要求从这批数据出发，估计常系数 α, β_1, β_2 的值。

这是一个回归分析问题，回归方程为 $Y = \alpha L^{\beta_1} K^{\beta_2} + \varepsilon$ ，显然，它不是线性回归方程，但是，如果我方程两边同时取对数，得到

$$\ln Y = \ln \alpha + \beta_1 \ln L + \beta_2 \ln K + \varepsilon^* \ ,$$

（原来有 $Y \approx \alpha L^{\beta_1} K^{\beta_2}$，误差项为 ε，取对数后有 $\ln Y \approx \ln \alpha + \beta_1 \ln L + \beta_2 \ln K$，也有一个误差项，我们个误差项记为 ε^*。）

在式子

$$\ln Y = \ln \alpha + \beta_1 \ln L + \beta_2 \ln K + \varepsilon^*$$

中，令 $y^* = \ln Y$， $\beta_0 = \ln \alpha$， $x_1 = \ln L$， $x_2 = \ln K$，它就化成了一个线性回归方程

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon^* \quad 。$$

用线性回归的方法可以求出它的解。

具体作回归时，所需要的观测数据 x_{i1} ， x_{i2}， y_i^* 用 $\ln L_i$ ， $\ln K_i$， $\ln Y_i$ 的数值代入，计算得到的

回归方程中常系数的估计 $\hat{\beta}_1$, $\hat{\beta}_2$，就是原来回归方程中 β_1, β_2 的估计，原来回归方程中 α 的估计，可以

$\hat{\alpha} = e^{\hat{\beta}_0}$ 求得。

例 7（1992 年全国数模竞赛 A 题）施肥效果分析

对 2 种作物——土豆、生菜，分别施以 3 种不同数量的肥料——氮、磷、钾，得到一批产量的数据，求施肥与产量之间的关系。

设 N, P, K 分别是氮、磷、钾肥的施肥量，Y 是产量。N, P, K 与 Y 之间，可能有各种各样的关系，这种关系显然不会是线性的。比如说，可以考虑下列关系：

$$Y = \beta_0 + \beta_1 N + \beta_2 P + \beta_3 K + \beta_4 N^2 + \beta_5 NP + \beta_6 NK + \beta_7 P^2 + \beta_8 PK + \beta_9 K^2 + \varepsilon \quad 。$$

这是一个 N, P, K 的 2 次多项式。

令 $x_1 = N$，$x_2 = P$，$x_3 = K$，$x_4 = N^2$，$x_5 = NP$，$x_6 = NK$，$x_7 = P^2$，$x_8 = PK$，$x_9 = K^2$，化成了一个线性回归方程

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_9 x_9 + \varepsilon \quad ，$$

可以用线性回归的方法求出它的解。

例 8 混合异辛烯催化反应

在混合异辛烯催化反应中，反应速度 y 与氢的分压 x_1 ，异辛烯的分压 x_2 ，异辛烷的分压 x_3 之间似有下列关系：

$$y=\frac{kx_1x_2}{(1+a\sqrt{x_1}+bx_2+cx_3)^3}+\varepsilon \ ,$$

其中，k,a,b,c 是常系数。现对 x_1,x_2,x_3,y 作观测，得到观测值 $(x_{i1},x_{i2},x_{i3},y_i)$，$i=1,2,\cdots,n$，要

系数 k,a,b,c 的估计值。

对回归方程两边开 3 次方，再取倒数，得到

$$\frac{1}{\sqrt[3]{y}}=\frac{1}{\sqrt[3]{k\,x_1x_2}}+\frac{a\sqrt{x_1}}{\sqrt[3]{k\,x_1x_2}}+\frac{b\,x_2}{\sqrt[3]{k\,x_1x_2}}+\frac{c\,x_3}{\sqrt[3]{k\,x_1x_2}}+\varepsilon^* \ ,$$

再令 $y^*=\frac{1}{\sqrt[3]{y}}$ ，$\beta_1=\frac{1}{\sqrt[3]{k}}$ ，$z_1=\frac{1}{\sqrt[3]{x_1x_2}}$ ，$\beta_2=\frac{a}{\sqrt[3]{k}}$ ，$z_2=\frac{\sqrt{x_1}}{\sqrt[3]{x_1x_2}}$ ，$\beta_3=\frac{b}{\sqrt[3]{k}}$ ，$z_3=\frac{x_2}{\sqrt[3]{x_1x_2}}$ ，$\beta_4=\frac{c}{\sqrt[3]{}}$

$z_4=\frac{x_3}{\sqrt[3]{x_1x_2}}$ ，原方程就化成了下列形式：

$$y^* = \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4 + \varepsilon \quad ,$$

这是一个不带常数项 β_0 的线性回归方程。对于这种回归方程，可以用求线性回归方程的解法，求得它的最小

乘解。作回归计算时，所需要的观测数据 z_{i1}，z_{i2}，z_{i3}，z_{i4}，y_i^*，用 $\frac{1}{\sqrt[3]{x_{i1}x_{i2}}}$，$\frac{\sqrt{x_{i1}}}{\sqrt[3]{x_{i1}x_{i2}}}$，$\frac{x_{i2}}{\sqrt[3]{x_{i1}x_{i2}}}$，$\frac{x_{i3}}{\sqrt[3]{x_{i1}x_{i2}}}$

$\frac{1}{\sqrt[3]{y_i}}$ 的数值代入，按线性回归方法求得常系数的估计 $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$ 后，从下列各式就可以求出原方程中各

数的估计值：

$$\hat{k} = \frac{1}{\hat{\beta}_1^3} \quad , \quad \hat{a} = \frac{\hat{\beta}_2}{\hat{\beta}_1} \quad , \quad \hat{b} = \frac{\hat{\beta}_3}{\hat{\beta}_1} \quad , \quad \hat{c} = \frac{\hat{\beta}_4}{\hat{\beta}_1} \quad 。$$

上面举了几个把非线性回归化为线性回归的例子。

一个非线性回归问题，如果能够象上面例子中所介绍的那样，通过适当的变量代换，化为线性回归，则称这回归为**广义线性回归**（Generalized Linear Regression）。

二、广义线性回归的一般形式和解法

设自变量 $x_1, x_2, \cdots, x_m$ 与因变量 y 之间，有下列关系：

$$y = f(\beta_0 + \beta_1 \varphi_1(x_1, \cdots, x_m) + \cdots + \beta_p \varphi_p(x_1, \cdots, x_m)) + \varepsilon \quad ,$$

其中，$y = f(y^*)$ 是已知的一元函数，有唯一的反函数 $y^* = f^{-1}(y)$ ，$\varphi_1(x_1, \cdots, x_m)$，$\varphi_2(x_1, \cdots, x_m)$，$\varphi_p(x_1, \cdots, x_m)$ 是自变量 $x_1, x_2, \cdots, x_m$ 的不含未知参数的函数，$\beta_0, \beta_1, \cdots, \beta_m$ 是常系数，$\varepsilon \sim N(0, \sigma^2$ 表示误差的随机变量，$\sigma > 0$ 。

对 $x_1, x_2, \cdots, x_m$， y 进行 n 次观测，得到观测值：

$$(x_{i1}, x_{i2}, \cdots, x_{im}, y_i) \quad , \quad i = 1, 2, \cdots, n \quad 。$$

求 $\beta_0, \beta_1, \cdots, \beta_m$ 的估计 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_m$，使得下式达到最小：

$$Q = \sum_{i=1}^{n} [\, y_i - f(\beta_0 + \beta_1 \varphi_1(x_{i1}, \cdots, x_{im}) + \cdots + \beta_p \varphi_p(x_{i1}, \cdots, x_{im}))]^2 \quad 。$$

这就是广义线性回归问题的一般形式。

对回归方程

$$y = f(\beta_0 + \beta_1 \varphi_1(x_1, \cdots, x_m) + \cdots + \beta_p \varphi_p(x_1, \cdots, x_m)) + \varepsilon$$

两边同时取反函数 f^{-1}，得到

$$f^{-1}(y) = \beta_0 + \beta_1 \varphi_1(x_1, \cdots, x_m) + \cdots + \beta_p \varphi_p(x_1, \cdots, x_m) + \varepsilon^* \quad 。$$

令 $y^* = f^{-1}(y)$，$z_1 = \varphi_1(x_1, \cdots, x_m)$，$\cdots$，$z_p = \varphi_p(x_1, \cdots, x_m)$，上述方程就化成了线性回归方程

$$y^* = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \varepsilon^* \quad 。$$

用线性回归的方法可以求出它的解。

三、广义线性回归中的加权处理

有些广义线性回归问题，化为线性时不需要取反函数 $y^* = f^{-1}(y)$，有些则要取反函数 $y^* = f^{-1}(y)$。要取反函数的广义线性回归问题，其实还有一点必须说明，就是：取了反函数后，得到的新问题并不完全等价于问题。

下面用简化的形式来说明这一点。

原问题 设自变量 x 与因变量 y 之间，有下列关系：

$$y = f(\beta_0 + \beta_1 x) + \varepsilon \quad 。$$

求 β_0, β_1 的估计 $\hat{\beta}_0, \hat{\beta}_1$，使得下式达到最小：$Q = \sum_{i=1}^{n} [y_i - f(\beta_0 + \beta_1 x_i)]^2$ 。

化为线性后的新问题 在自变量 x 与因变量 y 之间的关系式两边取反函数 $f^{-1}(y)$，得到

$$f^{-1}(y) = \beta_0 + \beta_1 x + \varepsilon^* \quad 。$$

求 β_0, β_1 的估计 $\hat{\beta}_0^*$ ，$\hat{\beta}_1^*$ ，使得下式达到最小：$Q^* = \sum_{i=1}^{n} [f^{-1}(y_i) - (\beta_0 + \beta_1 x_i)]^2$ 。

这两个问题不完全等价。因为变换 $f^{-1}(y)$ 把曲线变成直线，把原来各观测点到曲线的距离变成了各点到直线的距离。显然，原来各点到曲线的距离并不等于变换后各点到直线的距离，使各点到曲线的距离平方和 Q 最小的解，也不等于使各点到直线的距离平方和 Q^* 最小的解，所以 $\hat{\beta}_0^* \neq \hat{\beta}_0$，$\hat{\beta}_1^* \neq \hat{\beta}_1$ 。

为了解决这一问题，有人提出一种“加权处理”方法。

我们知道，当$a \approx b$时，有 $f'(a) \approx \dfrac{f(a)-f(b)}{a-b}$ ，即有 $f(a)-f(b) \approx f'(a)(a-b)$ 。

现在因为 $f^{-1}(y_i) \approx \beta_0 + \beta_1 x_i$ ，所以

$$y_i - f(\beta_0 + \beta_1 x_i) = f(f^{-1}(y_i)) - f(\beta_0 + \beta_1 x_i) \approx f'(f^{-1}(y_i))[f^{-1}(y_i) - (\beta_0 + \beta_1 x_i)] \text{ 。}$$

所以

$$Q = \sum_{i=1}^{n}[\, y_i - f(\, \beta_0 + \beta_1 x_i)\,]^2 \approx \sum_{i=1}^{n}\{f'(f^{-1}(y_i))[f^{-1}(y_i) - (\beta_0 + \beta_1 x_i)]\}^2 = \sum_{i=1}^{n} W_i[f^{-1}(y_i) - (\beta_0 + \beta_1 x_i)]$$

其中 $W_i = [f'(f^{-1}(y_i))]^2$ 称为**权**（Weight）。

因此，原问题可以近似等价于下列**加权回归问题**：

求 β_0, β_1 的估计 $\hat{\beta}_0^W, \hat{\beta}_1^W$ ，使得下式达到最小：

$$Q^W = \sum_{i=1}^{n} W_i[f^{-1}(y_i) - (\beta_0 + \beta_1 x_i)]^2 \text{ 。}$$

由于 $Q^W \approx Q$ ，所以求得的加权最小二乘估计 $\hat{\beta}_0^W \approx \hat{\beta}_0$ ， $\hat{\beta}_1^W \approx \hat{\beta}_1$ 。这也就是说，加权后得到的解，接近于原问题的解，比起不加权得到的解来，要好得多了。不过，加权毕竟是一种近似处理方法，加权后得到的也还不能说完全等价于原问题的解，这一点，也是要说明的。

四、广义线性回归应用的实例

例 9（2003 年华东地区数模竞赛题）向量场问题

有一个正方形的平面区域，在区域中的每一点 (x,y) 上，都定义了一个向量 (V_x,V_y) ，它们构成了一量场。现在已知其中 $n=81$ 个点上的向量数据，要求给出向量场的解析表达式。

V_x,V_y 都是坐标点 (x,y) 的函数，所以本题也就是要求两个函数表达式

$$V_x=V_x(x,y)，\ V_y=V_y(x,y) \ 。$$

题目中给出了两个数据表。**Table1** 中的数据点是整齐的网格点，向量值变化也很有规律，可以用插值方法 V_x,V_y 的表达式：

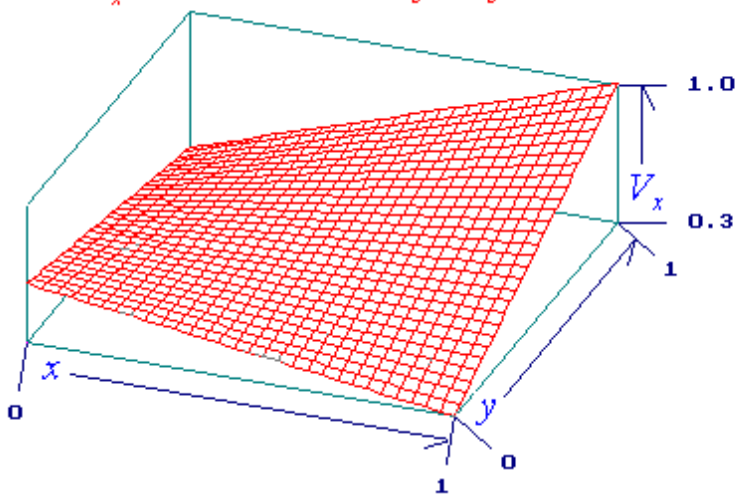
$$V_x=0.6-0.3x-0.3y+xy \ ，\ V_y=0.3-0.1x-0.15y+0.5xy \ 。$$

这是没有任何误差的精确表达式。

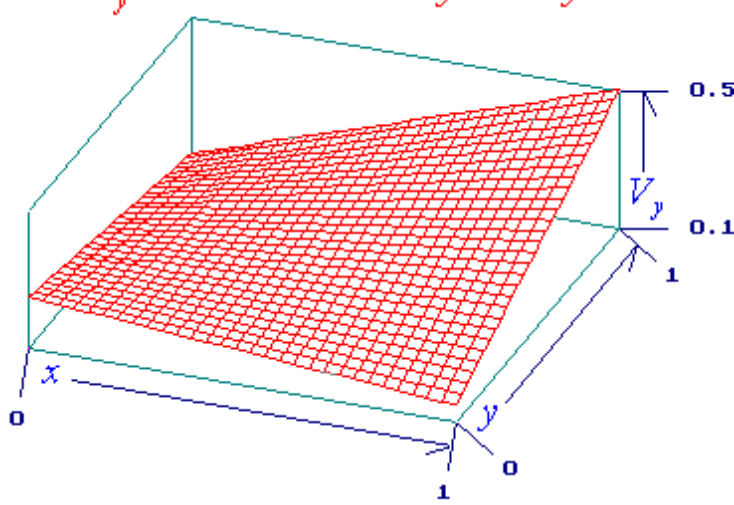
Table2 中的数据点是散乱的，不是整齐的网格点。

x_i	y_i	V_{xi}	V_{yi}
0.6554	0.4463	0.0129	0.0884
0.2010	0.8354	0.0220	−0.1464
0.8936	0.5264	0.0501	0.2104
⋮	⋮	⋮	⋮
0.7800	0.2820	−0.0401	0.0692

$V_x = 0.6 - 0.3x - 0.3y + xy$



$V_y = 0.3 - 0.1x - 0.15y + 0.5xy$



但是，从图像可以看出，Table2 数据的函数图像与 Table1 数据的函数图像十分相似，大致上是一个下列的双曲抛物面函数：

$$f(x, y) = a + bx + cy + dxy \quad 。$$

由于数据有一些误差，不能完全精确地满足上述方程，所以我们可以把它看作一个回归分析问题，V_x, V_y 的回归方程分别为：

$$V_x = a + bx + cy + dxy + \varepsilon \quad , \quad V_y = a' + b'x + c'y + d'xy + \varepsilon' \quad 。$$

其中 $a, b, c, d, a', b', c', d'$ 是待定未知常数。

令 $z_1 = x$， $z_2 = y$， $z_3 = xy$，就可以把它们化为线性回归方程：

$$V_x = a + bz_1 + cz_2 + dz_3 + \varepsilon \quad , \quad V_y = a' + b'z_1 + c'z_2 + d'z_3 + \varepsilon' \quad 。$$

用计算机软件可以很方便地求出问题的解。我们求得：

$$\hat{a} = 0.01999 \approx 0.02 , \quad \hat{b} = -0.19997 \approx -0.2 , \quad \hat{c} = -0.04996 \approx -0.05 , \quad \hat{d} = 0.49989 \approx 0.5 ,$$

$$\hat{a}' = 0.03998 \approx 0.04 , \quad \hat{b}' = -0.09997 \approx -0.1 , \quad \hat{c}' = -0.39999 \approx -0.4 , \quad \hat{d}' = 0.99996 \approx 1 \quad 。$$

所以 V_x, V_y 的表达式为：

$$V_x \approx 0.02 - 0.2x - 0.05y + 0.5xy = (x - 0.1)(y - 0.4)/2 ,$$

$$V_y \approx 0.04 - 0.1x - 0.4y + xy = (x - 0.4)(y - 0.1) \quad 。$$

例 10（1991 年国际数模竞赛 A 题）估计箱水流量

小镇上有一个高 40 英尺、底面直径 57 英尺的圆柱形水箱，每天向小镇上的居民供应生活用水。每隔一段测量一次水箱中的水位，测得数据如下（其中有两次水泵开动向水箱中加水的过程，在加水过程中没有水位记

编号	时刻（秒）	水位（0.01 英尺）
1	0	3175
2	3316	3110
3	6635	3054
4	10619	2994
5	13937	2947
6	17921	2892
7	21240	2850
8	25223	2795
9	28543	2752
10	32284	2697
11	35932	水泵开动
12	39332	水泵开动
13	39435	3550
14	43318	3445

编号	时刻（秒）	水位（0.01 英尺）
15	46636	3350
16	49953	3260
17	53936	3167
18	57254	3087
19	60574	3012
20	64554	2927
21	68535	2842
22	71854	2767
23	75021	2697
24	79254	水泵开动
25	82649	水泵开动
26	85968	3475
27	89953	3397
28	93270	3340

要求估计所有时刻（包括水泵开动期间）流出水箱的水流量 $f(t)$ 的变化情况，并估计一天的总用水量。

这个问题的最困难之处是有两次水泵开动加水的过程，加水过程中加了多少水，因为加水使水位变化了多
全都不知道。

为了简化问题，我们先不考虑加水过程，假定自始至终只有水流出，没有水加入。

设在时刻 t，水箱中水位为 $H = H(t)$，水箱中的总水量为 $\frac{\pi D^2}{4} H(t)$，其中，$D = 47$ 是水箱的底面直

在时刻 t，流出水箱的水流量

$$f(t) = -\frac{\mathrm{d}}{\mathrm{d}t}[\frac{\pi D^2}{4} H(t)] = -\frac{\pi D^2}{4}\frac{\mathrm{d}}{\mathrm{d}t} H(t) \quad 。$$

反过来，则有

$$\frac{\mathrm{d}}{\mathrm{d}t} H(t) = -\frac{4}{\pi D^2} f(t) \quad ,$$

$$H(t) = H_0 - \frac{4}{\pi D^2}\int_0^t f(t)\mathrm{d}t \quad 。$$

其中，$H_0 = H(0)$ 是 $t = 0$ 时的水位。

由于小镇上每天的用水情况周而复始，每天的流量变化几乎都是相同的，所以可以认为流量 $f(t)$ 是以
时间 $T=86400$ 秒为周期的周期函数。

我们知道，任何以 T 为周期的周期函数，都可以展开为下列形式的 Fourier 级数：

$$f(t)=a_0+a_1\cos(\frac{2\pi t}{T})+b_1\sin(\frac{2\pi t}{T})+a_2\cos(\frac{4\pi t}{T})+b_2\sin(\frac{4\pi t}{T})+\cdots \quad 。$$

作为近似，我们取级数前面的这 5 项。对它积分后可得：

$$H(t)=H_0-\frac{4}{\pi D^2}\int_0^t f(t)\mathrm{d}t$$

$$=\beta_0+\beta_1 t+\beta_2\sin(\frac{2\pi t}{T})+\beta_3\cos(\frac{2\pi t}{T})+\beta_4\sin(\frac{4\pi t}{T})+\beta_5\cos(\frac{4\pi t}{T}) \quad 。$$

这是一个可以化为线性的广义线性回归方程。

令 $x_1=t$，$x_2=\sin(\frac{2\pi t}{T})$，$x_3=\cos(\frac{2\pi t}{T})$，$x_4=\sin(\frac{4\pi t}{T})$，$x_5=\cos(\frac{4\pi t}{T})$，上式就化为

$$H=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\varepsilon \quad ,$$

我们就得到了一个多元线性回归方程。

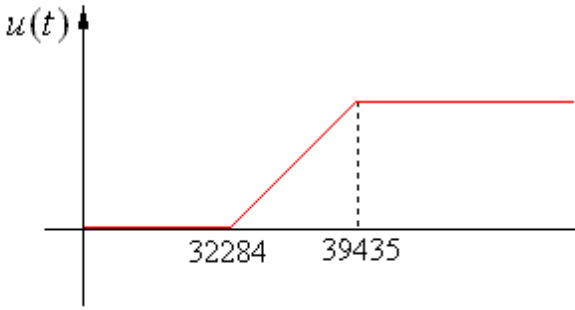
用多元线性回归方法，从水位 H 和时间 t 的观测数据出发，可以直接求出它的解。

以上是在不考虑加水过程的情况下得到的结果。下面我们进一步把加水过程考虑进去。

设加水速度是均匀的，第一次加水和第二次加水过程中，水位上升的速度分别是两个未知常数 β_6 和 β

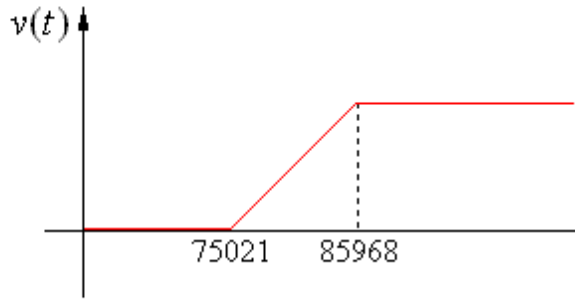
第一次加水，相当于在水位函数式上加上一项 $\beta_6 u(t)$，其中

$$u(t)=\begin{cases}0 & t<32284\\ t-32284 & 32284\le t<39435\\ 39435-32284 & t\ge 39435\end{cases}。$$



第二次加水，相当于在水位函数式上加上一项 $\beta_7 v(t)$，其中

$$v(t)=\begin{cases}0 & t<75021\\ t-75021 & 75021\le t<85968\\ 85968-75021 & t\ge 85968\end{cases}。$$



把两次加水过程考虑进去，回归方程成为

$$H=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6u(t)+\beta_7v(t)+\varepsilon\ 。$$

令 $x_6 = u(t)$， $x_7 = v(t)$， 上式就化为线性回归方程

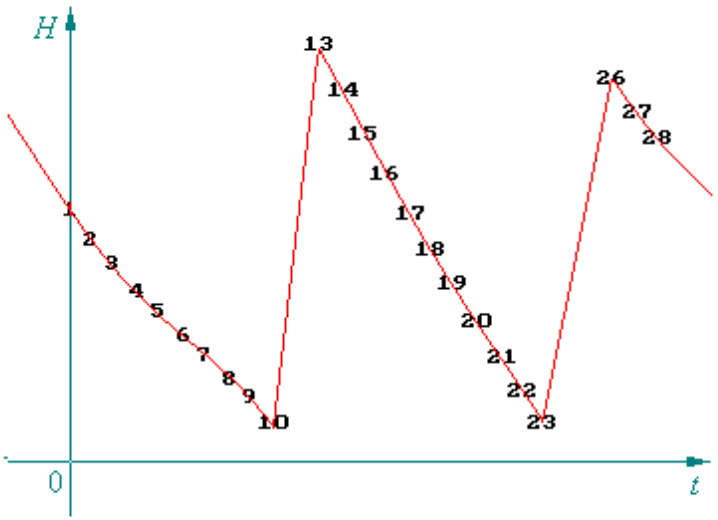
$$H = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon \quad 。$$

用计算机软件可以求得未知参数的估计

$$\hat{\beta}_0 = 3242.99，\ \hat{\beta}_1 = -0.0201157，\ \hat{\beta}_2 = 38.4813，\ \hat{\beta}_3 = -62.2651，$$

$$\hat{\beta}_4 = -22.8893，\ \hat{\beta}_5 = -3.23168，\ \hat{\beta}_6 = 0.141780，\ \hat{\beta}_7 = 0.0926685。$$

把这些值代入回归方程，作出水位 H 随时间 t 变化的函数图像如下：



从图像可以看出，用广义线性回归求出的回归方程函数曲线与观测值点符合得很好，回归效果是十分令人的。

表达式 $H=\hat{\beta}_0+\hat{\beta}_1 t+\hat{\beta}_2 \sin(\frac{2\pi t}{T})+\hat{\beta}_3 \cos(\frac{2\pi t}{T})+\hat{\beta}_4 \sin(\frac{4\pi t}{T})+\hat{\beta}_5 \cos(\frac{4\pi t}{T})+\hat{\beta}_6 u(t)+\hat{\beta}_7 v(t)$ 水位 H 的变化，实际上包括两部分：一部分是由于小镇用水引起的水位的下降，一部分是由于水泵加水引水位的上升。

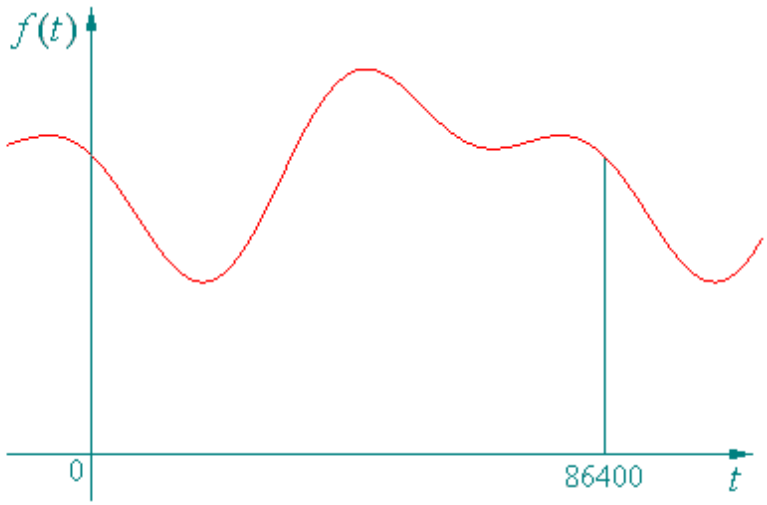
从回归方程中去掉代表加水过程的两项：$\hat{\beta}_6 u(t)$ 和 $\hat{\beta}_7 v(t)$，就得到了不考虑加水过程的情况下纯粹因镇用水引起水位 H 随时间 t 变化的函数表达式：

$$H(t)=\hat{\beta}_0+\hat{\beta}_1 t+\hat{\beta}_2 \sin(\frac{2\pi t}{T})+\hat{\beta}_3 \cos(\frac{2\pi t}{T})+\hat{\beta}_4 \sin(\frac{4\pi t}{T})+\hat{\beta}_5 \cos(\frac{4\pi t}{T}) \quad 。$$

再对它求导，就得到在不考虑加水过程的情况下流量 $f(t)$ 随时间 t 变化的函数式：

$$f(t)=-\frac{\pi D^2}{4}\frac{\mathrm{d}}{\mathrm{d}t}H(t)=-\frac{\pi^2 D^2}{2T}\left[\hat{\beta}_1\frac{T}{2\pi}+\hat{\beta}_2\cos(\frac{2\pi t}{T})-\hat{\beta}_3\sin(\frac{2\pi t}{T})+2\hat{\beta}_4\cos(\frac{4\pi t}{T})-2\hat{\beta}_5\sin(\frac{4\pi t}{T})\right]$$

函数图像为



要求出一天 24 小时即 $T=86400$ 秒的总用水量，可以将 $t=0$ 和 $t=T$ 代入（去掉加水过程的）$H(t)$ 的函数式，求出一天开始时的水位和一天终了时的水位

$$H(0)=\hat{\beta}_0+\hat{\beta}_1 0+\hat{\beta}_2\sin(0)+\hat{\beta}_3\cos(0)+\hat{\beta}_4\sin(0)+\hat{\beta}_5\cos(0)=\hat{\beta}_0+\hat{\beta}_3+\hat{\beta}_5 \ ,$$

$$H(T)=\hat{\beta}_0+\hat{\beta}_1 T+\hat{\beta}_2\sin(2\pi)+\hat{\beta}_3\cos(2\pi)+\hat{\beta}_4\sin(2\pi)+\hat{\beta}_5\cos(2\pi)=\hat{\beta}_0+\hat{\beta}_1 T+\hat{\beta}_3+\hat{\beta}_5 \ ,$$

水位之差是

$$H(0)-H(T)=-\hat{\beta}_1 T \ 。$$

得到这么一个简单的结果，其实并不奇怪。因为，在水位 $H(t)$ 的函数表达式中，除了 $\hat{\beta}_1 t$ 以外，其他各是以一天时间 T 为周期的周期项，经过一个周期 T 以后，这些项的值又返回到了初始状态，前后一减，自消去了。

然后，用下式就可以计算出一天的总用水量：

$$\frac{\pi D^2}{4}[H(0)-H(T)]=-\frac{\pi D^2}{4}\hat{\beta}_1 T=\frac{3.141593\times 57^2}{4}\times 0.02011572\times 0.01\times 86400$$

$$=44349.5\text{（英尺}^3\text{）}=331756\text{（加仑）}=1255.84\text{（米}^3\text{）} \ 。$$

§5 拟线性回归（Quasi-linear Regression）

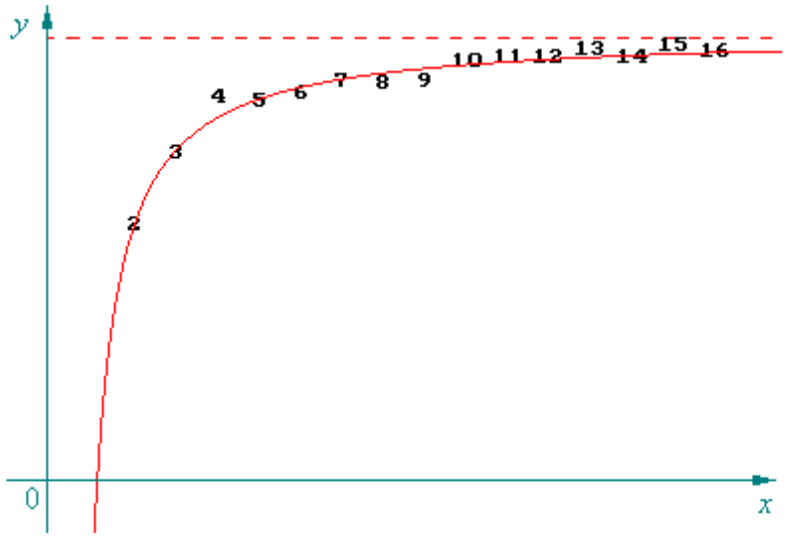
例 11 炼钢炉的炉龄与产量

对炼钢炉的炉龄 x 和产量 y 进行观测，得到下列数据：

x	2	3	4	5	6	7	8	9
y	6.42	8.20	9.58	9.50	9.70	10.00	9.93	9.99

x	10	11	12	13	14	15	16
y	10.49	10.59	10.60	10.80	10.60	10.90	10.76

从图像可以看出，产量 y 随着炉龄 x 的增大而增大，但它不会无限增大下去，而是逐渐趋于一个极限，以，可以认为，x 与 y 之间有下列关系：$y = a + \dfrac{b}{x+c} + \varepsilon$ ，其中 a,b,c 是未知常数。



这个回归方程，虽然与线性回归方程很相似，但是却无法用变量代换化为线性。

类似的情形还有很多，例如：

$$y = a + b\ln(x + c) + \varepsilon \quad , \quad y = a + b\mathrm{e}^{cx} + \varepsilon \quad , \quad y = a + bx^{c} + \varepsilon \quad ,$$

等等，像这样的回归，就称为**拟线性回归**。

拟线性回归的一般形式为：设自变量 x 与因变量 y 之间，有下列关系：

$$y = a + b\,f(x;c) + \varepsilon \quad ,$$

其中 a,b,c 是未知常数，f 是形式已知的函数，含有一个未知参数 c，$\varepsilon \sim N(0,\sigma^2)$ 是表示误差的随机变

$\sigma > 0$ 。

对 x，y 进行 n 次观测，得到观测值

$$(x_i\ ,\ y_i) \quad , \quad i = 1, 2, \cdots, n \quad 。$$

求 a,b,c 的估计 $\hat{a},\hat{b},\hat{c}$，使得下式达到最小：

$$Q = \sum_{i=1}^{n}[\ y_i - a - bf(x_i;c)\ ]^2 \quad 。$$

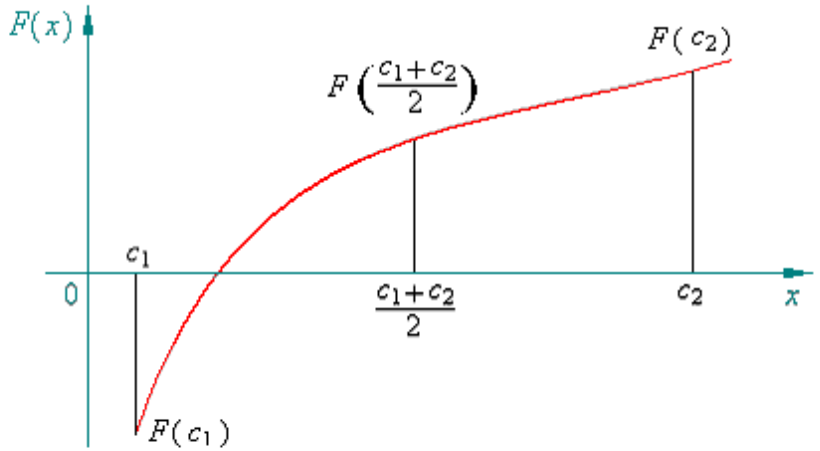
Q 是 a,b,c 的函数，所以，这是一个多元函数求最小值的问题，我们可以通过求偏导数、解下列方程组

方法，来确定 Q 的最小值点：

$$\begin{cases} \dfrac{\partial Q}{\partial a} = 0 \\ \dfrac{\partial Q}{\partial b} = 0 \\ \dfrac{\partial Q}{\partial c} = 0 \end{cases}$$

从这个方程组中消去 a,b，得到一个只含有 c 的方程 $F(c)=0$。

$F(c)=0$ 是一个复杂的一元方程，无法用解析方法求精确解，但是可以用数值方法求近似解。例如可以用分法”求解。



先用试探的方法找到两个值 c_1 和 c_2，使得 $F(c_1)$ 和 $F(c_2)$ 的符号恰好是一正一负，然后看区间 $[c_1,$

中点上的函数值 $F(\frac{c_1+c_2}{2})$：如果 $F(\frac{c_1+c_2}{2})$ 与 $F(c_1)$ 正负异号，则将搜索区间缩小为 $[c_1,\frac{c_1+c_2}{2}]$；

果 $F(\frac{c_1+c_2}{2})$ 与 $F(c_2)$ 正负异号，则将搜索区间缩小为 $[\frac{c_1+c_2}{2},c_2]$。就这样，每次将搜索区间缩小一

直到 $|c_1-c_2|<\varepsilon$ 为止（ε 是事先给定的误差水平界限）。

用数值方法求出 c 的估计值以后，将它代入方程组 $\begin{cases}\frac{\partial Q}{\partial a}=0\\ \frac{\partial Q}{\partial b}=0\\ \frac{\partial Q}{\partial c}=0\end{cases}$，就可以求出 a,b 的估计值。

例如，前面的炼钢炉的炉龄与产量的例子，用上述方法，借助计算机软件可以求得

$$\hat{a}=11.2154，\ \hat{b}=-7.57050，\ \hat{c}=-0.414421。$$

§6 非线性回归（Non-linear Regression）

一、非线性回归问题的一般形式和解法

设自变量 $x_1, x_2, \cdots, x_m$ 与因变量 y 之间有下列关系：

$$y = F(x_1, x_2, \cdots, x_m; a_1, a_2, \cdots, a_p) + \varepsilon \quad ,$$

其中，F 是形式已知的非线性函数，$a_1, a_2, \cdots, a_p$ 是函数中的未知参数。

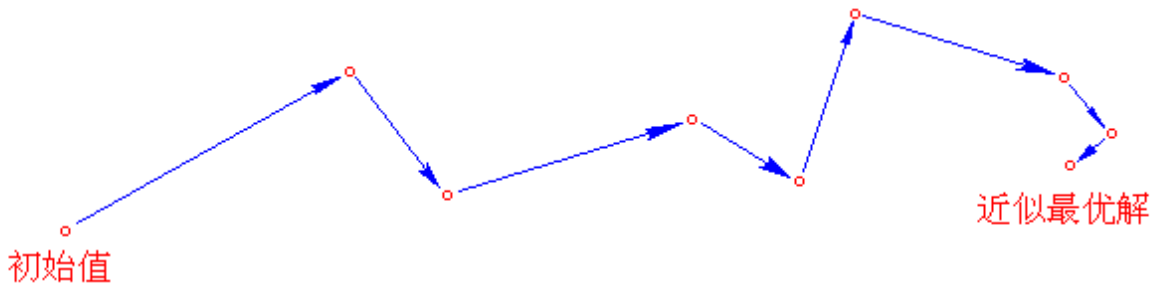
进行 n 次观测，得到观测值 $(x_{i1}, x_{i2}, \cdots, x_{im}, y_i)$ ，$i = 1, 2, \cdots, n$ 。

求 $a_1, a_2, \cdots, a_p$ 的估计 $\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p$ ，使得下列平方和 Q 达到最小：

$$Q = \sum_{i=1}^{n} [\, y_i - F(x_{i1}, x_{i2}, \cdots, x_{im}; a_1, a_2, \cdots, a_p \,)]^2 \quad 。$$

Q 是 $a_1, a_2, \cdots, a_p$ 的函数，所以，这是一个多元函数求最小值的问题。但是，由于非线性函数过于复

一般来说，我们不可能通过求偏导数、解方程组 $\begin{cases} \dfrac{\partial Q}{\partial a_1} = 0 \\ \dfrac{\partial Q}{\partial a_2} = 0 \\ \cdots\cdots \\ \dfrac{\partial Q}{\partial a_p} = 0 \end{cases}$ 的方法，来确定 Q 的最小值点。



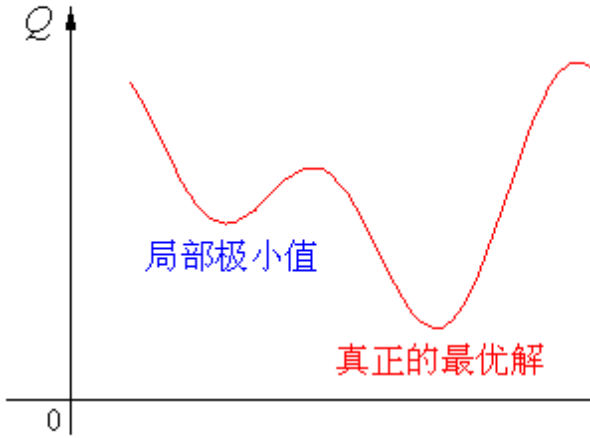
其实，对于这样一个多元函数求最小值的问题，人们已经提出了许多求近似解的数值方法。这些方法的思想是：从一个初始值出发，逐步搜索最优解，搜索的步长逐渐缩小，当搜索的步长或最优解的变化小于事先的一个误差水平界限时，搜索结束，给出问题的近似解。

在借助计算机软件进行求解的过程中，可能会出现许多问题，例如：

（1）逐步搜索超出函数的定义域，发生计算溢出；

（2）发生“死循环”，长期得不到解；

（3）得到的解不是真正的最优解，等等。



要避免发生这样的问题，选择适当的初始值非常重要。最好能够根据问题的实际意义，参考文献资料和类似先例，确定比较合理的初始值。还有一种方法，是将问题先作一些简化或转换，变成能用其他回归方法求解的问题，用其他方法求出初步的近似解，再用这个解作为非线性回归的初始值。

二、非线性回归应用的实例

例 12（1990 年上海市数模竞赛 A 题）脑血流量测定

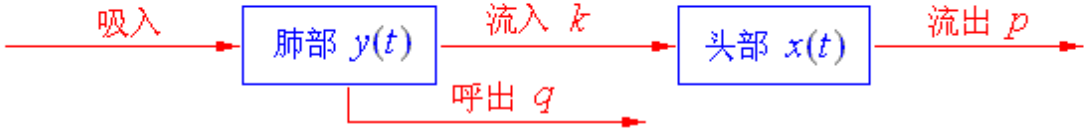
为了测定脑血流量，让受试者吸入含有放射性同位素的气体。放射性同位素随血流进入头部，又由脑血流带同时，吸入肺部的放射性同位素由呼出气体带出。

对头部的放射性计数率和呼出气的放射性计数率同时进行测试，测得一批数据如下：

编号	时间	头部计数率	呼出气计数率
1	1.00	1534	2231
2	1.25	1528	1534
3	1.50	1468	1054
4	1.75	1378	724
5	2.00	1272	498
6	2.25	1162	342
7	2.50	1052	235
8	2.75	947	162
9	3.00	848	111
10	3.25	757	76
11	3.50	674	52
12	3.75	599	36
13	4.00	531	25
14	4.25	471	17
15	4.50	417	12
16	4.75	369	8
17	5.00	326	6
18	5.25	288	4

编号	时间	头部计数率	呼出气计数率
20	5.75	225	2
21	6.00	199	1
22	6.25	175	1
23	6.50	155	1
24	6.75	137	1
25	7.00	121	0
26	7.25	107	0
27	7.50	94	0
28	7.75	83	0
29	8.00	73	0
30	8.25	65	0
31	8.50	57	0
32	8.75	50	0
33	9.00	44	0
34	9.25	39	0
35	9.50	35	0
36	9.75	31	0
37	10.0	27	0

已知头部计数率下降的速率与当时头部的计数率成正比，比例系数称为**脑血流量系数**。头部计数率上升的速率与当时呼出气的计数率成正比。要求建立数学模型，求出脑血流量系数。



设 $x(t)$ 是时刻 t 时头部的计数率，$y(t)$ 是时刻 t 时呼出气的计数率。

根据题意，有（其中 k, p 是常数，p 就是脑血流量系数）

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = ky(t) - px(t) \quad 。$$

同时，根据放射性衰变原理，呼出气的放射性计数率减少的速率与当时呼出气的计数率成正比，即有（其中 是常数）

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = -qy(t) \quad 。$$

因此，有联立微分方程组：

$$\begin{cases} \dfrac{\mathrm{d}x(t)}{\mathrm{d}t} = ky(t) - px(t) \\ \dfrac{\mathrm{d}y(t)}{\mathrm{d}t} = -qy(t) \end{cases}$$ ，初始条件 $$\begin{cases} x(0) = x_0 \\ y(0) = y_0 \end{cases}$$ 。

解这个微分方程组，可得：

$$\begin{cases} x(t) = \dfrac{ky_0}{p-q}\mathrm{e}^{-qt} + (x_0 - \dfrac{ky_0}{p-q})\mathrm{e}^{-pt} = C_1\mathrm{e}^{-qt} + C_2\mathrm{e}^{-pt} \\ y(t) = y_0\mathrm{e}^{-qt} \end{cases} 。$$

$x(t)$， $y(t)$ 的表达式都是非线性的。

对于 $y(t) = y_0\mathrm{e}^{-qt}$ ，可以两边取对数，将它化为线性形式 $\ln y(t) = \ln y_0 - qt$ 。

化为线性后，用线性回归的方法求解，解得 $\hat{y}_0 \approx 8373.5$ ， $\hat{q} \approx 4.5243$ 。

因为在将广义线性方程化为线性时，对因变量 y 作了变换，所以，上面得到的解还不是原问题的真正的

用这个解作为初始值，进行非线性回归，求得原问题的真正的解为 $\hat{y}_0 = 10000.3 \approx 10000$ ， $\hat{q} = 1.50005 \approx 1$

将它们代入 $x(t)$ 的表达式，有

$$x(t) = C_1\mathrm{e}^{-1.5t} + C_2\mathrm{e}^{-pt} 。$$

这是一个不能化为线性的非线性函数，对它进行非线性回归，最后得到

$$\hat{C}_1 = -3995.90 \approx -4000 ， \hat{C}_2 = 3998.91 \approx 4000 ， \hat{p} = 0.499925 \approx 0.5 。$$

脑血流量系数 p 为 0.5 。

例 13　肉鸡饲养问题

已知每单位重量肉鸡的价格是 a（单位：元/kg），每单位重量肉鸡每天的饲料费是 b（单位：元/（kg 日
每只鸡每天的管理费是 c（单位：元/日），研究肉鸡饲养到什么时候出售得到的利润最大。

要解决这一问题，首先要知道肉鸡饲养天数与肉鸡重量的函数关系。对肉鸡的饲养天数 t（单位：日）
鸡的重量 $y(t)$（单位：kg）进行观测，得到一组数据如下：

饲养天数	4	8	12	16	20	24	28	32	36
重量	0.070	0.119	0.198	0.297	0.434	0.606	0.803	1.027	1.245

饲养天数	40	44	48	52	56	60	64	68
重量	1.488	1.736	1.980	2.170	2.450	2.687	2.915	3.095

可以认为，肉鸡重量 $y(t)$ 与饲养天数 t 之间的关系，满足下列微分方程及初始条件：

$$\begin{cases} \dfrac{\mathrm{d}y(t)}{\mathrm{d}t} = k\ y(t)\left[1 - \dfrac{y(t)}{w}\right] \\ y(0) = y_0 \end{cases}$$

其中，k, w, y_0 是未知常数。

在初始条件下解微分方程，可求得

$$y(t)=\frac{w}{1+(\frac{w}{y_0}-1)\mathrm{e}^{-kt}} \quad 。$$

这样的函数曲线称为 logistic **曲线**。在上式中，常数 w, y_0, k 未知，所以，要得到肉鸡重量 $y(t)$ 与饲数 t 之间的关系式，必须从 t 与 $y(t)$ 的观测数据出发，求出 w, y_0, k 的估计值。

这是一个非线性回归问题，为了在逐步搜索时有一个比较合理的初始值，可以先对原回归方程两边取倒数到

$$\frac{1}{y(t)}=\frac{1}{w}+\left(\frac{1}{y_0}-\frac{1}{w}\right)\mathrm{e}^{-kt} \quad 。$$

令 $y^*=\frac{1}{y(t)}$ ，$a=\frac{1}{w}$，$b=\frac{1}{y_0}-\frac{1}{w}$，$c=-k$，上式化为

$$y^*=a+b\mathrm{e}^{ct} \quad 。$$

这是一个拟线性回归方程，可以用求拟线性回归的方法求解，求得

$$\hat{a}=0.48442，\ \hat{b}=23.1385，\ \hat{c}=-0.131454。$$

即有

$$\hat{w}=\frac{1}{\hat{a}}=2.0643 \quad ，\quad \hat{y}_0=\frac{1}{\hat{a}+\hat{b}}=0.04233 \quad ，\quad \hat{k}=-\hat{c}=0.131454 \quad 。$$

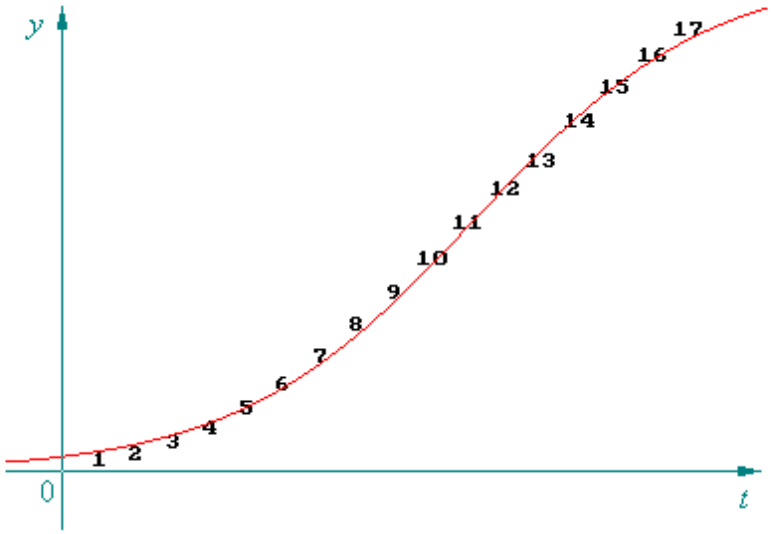
由于在把非线性回归转换成拟线性回归的过程中，对因变量作了代换，所以，上面得到的解，还不是真正原题的解。用上述解作为初始值，借助计算机软件，对原问题作非线性回归，计算求得 w, y_0, k 的更精确的估计

$$\hat{w} = 3.53348 \ , \ \hat{y}_0 = 0.106883 \ , \ \hat{k} = 0.0776665 \ 。$$

这样我们就得到了肉鸡重量 $y(t)$ 与饲养天数 t 之间的函数关系

$$y(t) = \frac{3.53348}{1+(\frac{3.53348}{0.106883}-1)\mathrm{e}^{-0.0776665t}} = \frac{3.53348}{1+32.0593\ \mathrm{e}^{-0.0776665t}} \ 。$$

它的函数图像如下：



下面进一步考虑什么时候出售肉鸡获利最多。

根据已知条件，一只鸡饲养 t 天后出售，获得净利润为 $z(t)=a[y(t)-y_0]-b\sum_{i=1}^{t}y(i)-ct$ 。

我们的目标是求 t 的值，使得利润 $z(t)$ 达到最大。这是一个一元函数求最大值的问题，可以用求导，令

数等于 0，解方程 $\frac{\mathrm{d}z(t)}{\mathrm{d}t}=0$ 的方法来求解。

为了便于求导，把目标函数式中的求和变成积分。我们知道 $\sum_{i=1}^{t}y(i)\approx\int_0^t y(t)\mathrm{d}t$ 。所以

$$z(t)=a[y(t)-y_0]-b\sum_{i=1}^{t}y(i)-ct\approx a[y(t)-y_0]-b\int_0^t y(t)\mathrm{d}t-ct \quad 。$$

$$\frac{\mathrm{d}z(t)}{\mathrm{d}t}=a\frac{\mathrm{d}y(t)}{\mathrm{d}t}-by(t)-c=akr\,y(t)\left[1-\frac{y(t)}{w}\right]-by(t)-c=0 \quad 。$$

整理后有

$$[y(t)]^2+\left(\frac{bw}{akr}-w\right)y(t)+\frac{cw}{akr}=0 \quad 。$$

这是一个关于 $y(t)$ 的一元二次代数方程，用求根公式可求得两个解：

$$y(t)=y_1 \quad , \quad y(t)=y_2 \quad 。$$

将这两个解代入 $z(t)$ 的二阶导数表达式

$$\frac{\mathrm{d}^2 z(t)}{\mathrm{d}t^2} = [ak - \frac{2ak}{w} y(t) - b]\, y(t)\, [1 - \frac{y(t)}{w}] \quad ,$$

可以发现 $\left.\frac{\mathrm{d}^2 z(t)}{\mathrm{d}t^2}\right|_{y=y_1} < 0$ ， $\left.\frac{\mathrm{d}^2 z(t)}{\mathrm{d}t^2}\right|_{y=y_2} > 0$ ，说明 $y(t) = y_1$ 使 $z(t)$ 达到最大， $y(t) = y_2$ 使使 $z(t)$ 达到

最小。当然应该取 $y(t) = y_1$ 。

还要求 t 取什么值时，才有 $y(t) = y_1$ 使得 $z(t)$ 达到最大。

从

$$\frac{w}{1 + (\frac{w}{y_0} - 1)\mathrm{e}^{-kt}} = y(t) = y_1$$

可以解出最佳出售时刻为

$$t = \frac{1}{k} \ln \left(\frac{\frac{w}{y_0} - 1}{\frac{w}{y_1} - 1} \right) = t_1 \quad 。$$

将 $t = t_1$ 代入 $z(t)$，还可以求出按最佳时刻出售获得利润的最大值。

§7　微分方程中未知参数的估计

前面的两个例子“脑血流量测定”和“肉鸡饲养问题”，都是先解微分方程，求得函数表达式，然后用回析求出表达式中未知参数的估计。

但是，在许多实际问题中，微分方程往往很复杂，无法求出精确的解析函数表达式解，只能求出近似的数值对于这样的问题，需要从观测数据出发，直接求微分方程中未知参数的估计。

设 x 是自变量，$y_1, y_2, \cdots, y_m$ 是因变量，它们满足下列微分方程组：

$$\begin{cases} \dfrac{\mathrm{d}y_1}{\mathrm{d}x} = F_1(x, y_1, y_2, \cdots, y_m; a_1, a_2, \cdots, a_p) \\ \qquad\qquad \vdots \\ \dfrac{\mathrm{d}y_m}{\mathrm{d}x} = F_m(x, y_1, y_2, \cdots, y_m; a_1, a_2, \cdots, a_p) \end{cases}$$

其中，$a_1, a_2, \cdots, a_p$ 是未知参数。

如果 $a_1, a_2, \cdots, a_p$ 的值已经给定，同时已知微分方程组的初始条件，就可以用数值方法（例如 Runge-K法，Hamming 法，等等）求出微分方程组的解：

$$\begin{cases} y_1 = y_1(x; a_1, a_2, \cdots, a_p) \\ \qquad\qquad \vdots \\ y_m = y_m(x; a_1, a_2, \cdots, a_p) \end{cases}。$$

但是，实际上 $a_1,a_2,\cdots,a_p$ 是未知参数，需要从观测数据出发进行估计。

设对自变量和因变量进行了 n 次观测，得到观测值 $(x_i,y_{i1},y_{i2},\cdots,y_{im})$ ，$i=1,2,\cdots,n$ 。

我们的目标是：要求 $a_1,a_2,\cdots,a_p$ 的估计，使得下列加权平方和达到最小：

$$Q=\sum_{j=1}^{m}W_j^2\sum_{i=1}^{n}[\,y_{ij}-y_j(x_i;a_1,a_2,\cdots,a_p)\,]^2 \quad 。$$

其中 W_j 是因变量 y_j 的加权系数。因为问题中有多个因变量，每个因变量的实际意义不一样，量纲单小不一样，对精度的要求也不一样，所以，在求最小二乘估计时，必须对各个因变量加上不同的权，变成求加方和最小值的问题。加权系数 W_j 越大，y_j 的估计越精确；加权系数 W_j 越小，y_j 的估计越不精确。一$W_j=\dfrac{1}{S_j}$，其中 S_j 是 y_j 的样本标准差。

求微分方程中未知参数的估计，类似于作非线性回归：从 $a_1,a_2,\cdots,a_p$ 的一组初始值出发，通过逐步搜求 Q 的最小值点。与非线性回归不同的是，每搜索一步，就要根据当前的 $a_1,a_2,\cdots,a_p$ 的值，用数值方法次微分方程。显然，求微分方程中未知参数的估计，比起非线性回归来，计算速度要慢得多，而且初始值的适取，显得更加重要。

例 14 一个求微分方程未知参数估计的例子

设变量 x, y_1, y_2 满足下列微分方程组

$$\begin{cases} \dfrac{dy_1}{dx} = a(x+1)y_1 + \sin x \\ \dfrac{dy_2}{dx} = \dfrac{by_1}{x+1} \end{cases} 。$$

有观测数据

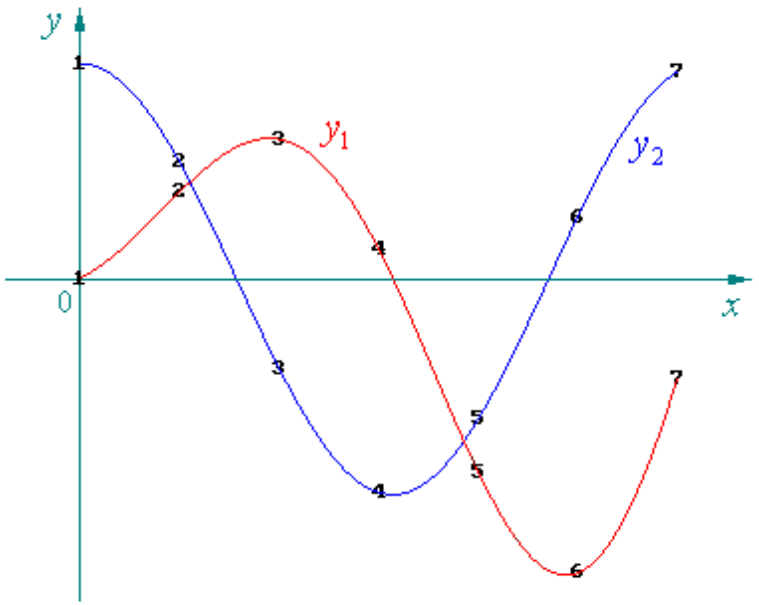
x	0	1	2	3	4	5	6
y_1	0.0000	1.6829	2.7279	0.5645	-3.7840	-5.7535	-1.9559
y_2	2.0000	1.0806	-0.8323	-1.9800	-1.3073	0.5673	1.9203

求微分方程中未知参数 a,b 的估计。

按照前面介绍的解法，用计算机软件求解，求解中，取 y_1 的加权系数 $W_1 = \dfrac{1}{7.46792}$，取 y_2 的加权系数 $W_2 = \dfrac{1}{3.89865}$，最后求得 a,b 的估计为

$$\hat{a} = 0.499984\ ,\ \hat{b} = -2.00136\ 。$$

下面是用估计值代入得到的 y_1 和 y_2 的函数图像：



从图像可以看出，函数曲线与观测值点符合得很好，参数估计的结果还是比较令人满意的。